An Evaluation of Factors Influencing Bayesian Learning Systems

Eric L. Eisenstein, M.B.A. and Farrokh Alemi, Ph.D.

Health Administration Program
Cleveland State University
Cleveland, Ohio 44115

ABSTRACT

This paper examines the influences of situational and model factors upon the accuracy of Bayesian learning systems. In particular, it is concerned with the impact of variations in training sample size, number of attributes, choice of Bayesian model, and criteria for excluding model attributes upon the overall accuracy of the simple and proper Bayes models.

INTRODUCTION

"Conditional independence means that the presence of one clue does not change the value of any other clue."[1] During the past two decades, there has been considerable discussion regarding the importance of the conditional independence assumption in Bayesian analysis. One group of researchers has stated that the problems encountered in managing conditional nonindependence have been an impediment to the acceptance of Bayesian analysis by the medical and other potentially interested communities.[2, 3, 4, 5] In contrast, other researchers have asserted that the simple Bayesian model is relatively robust and that the management of conditional nonindependence is not an important problem.[6, 7] This study reviewed factors which have been assumed to influence the importance of the conditional independence assumption in Bayesian analysis. It also identified situations in which Bayesian models that assume conditional independence and conditional nonindependence in their data are respectively most appropriate.

BAYESIAN TERMINOLOGY

Bayesian analysis is often concerned with predicting whether one of two mutually exclusive and collectively exhaustive events will occur. The probability of such events occurring, without considering the presence of additional data, is called the prior probability. The probability of such events occurring, given the presence of additional data, is called the posterior probability. Bayes first suggested a formula for computing the posterior odds of an event.[8] This formula is given below.

$$\textit{Posterior Odds of } H = [p(D_1,D_2,...,D_n|H_1)/p(D_1,D_2,...,D_n|H_0)]^* [p(H_1)/p(H_0)] \tag{1}$$

Formula (1) states that the posterior odds of event H are equal to the product of two terms. The first term is the likelihood ratios for each attribute in the set of additional data which are used to predict event H. The second term is the prior odds of event H before the effect of the additional set of data is considered. The individual likelihood ratios in formula (1) can be expanded as shown below. In this formula, the data are assumed to be conditionally independent.

$$p(D_1,D_2,...,D_n|H_1)/p(D_1,D_2,...,D_n|H_0) = [p(D_1|H_1)/p(D_1|H_0)]^*[p(D_2|H_1)/p(D_2|H_0)]^*...^*[p(D_n|H_1)/p(D_n|H_0)] \tag{2}$$

Formula (2) above for computing likelihood ratios assumes that the individual attributes in the data set are conditionally independent. If this assumption is not valid and the attributes are interdependent, this formula must be modified to account for the joint likelihood ratios in the set of attributes. This revised formula is:

$$p(D_1,D_2,...,D_n|H_1)/p(D_1,D_2,...,D_n|H_0) = [p(D_1|H_1)/p(D_1|H_0)]^* [p(D_2|H_1,D_1)/p(D_2|H_1,D_1)]^* ...^* [p(D_n|H_1,D_1,D_2,...,D_{n-1})/p(D_n|H_0,D_1,D_2,...,D_{n-1})] \tag{3}$$

Each term in this formula is conditioned on the values for all previous terms. Only the first term is in its conditionally independent form. All other terms are in their conditionally nonindependent forms. Formulas (2) and (3) above are the primary Bayesian models which have been used in previous studies. We will call formula (2) the simple Bayes model and formula (3) the proper Bayes model.

0195-4210/92/$5.00 © 1994 AMIA, Inc.

SITUATIONAL FACTORS

Situational factors influence the relative efficacy of the simple and proper Bayes models. These factors and their interactions are often assumed to determine which of the Bayes models will have the greatest accuracy.

(1) Conditional Nonindependence:

The existence of conditional nonindependence in training data sets has a direct bearing upon the choice between the simple and proper Bayes models. If data are conditionally independent, there are no joint likelihood ratios and formula (2) can used for computations. With formula (2), each attribute's likelihood ratio is computed from the entire training sample. In contrast, when data are conditionally nonindependent, joint likelihood ratios exist and formula (3) should be used to calculate the conditional likelihood ratios. These conditional likelihood ratios are computed by successively reducing the training sample to account for dependencies with attributes which were previously included in the model.

(2) Training Sample Size:

Training sample size is generally considered to be a primary determinant of accuracy in both the simple and the proper Bayes models. Chard generated conditionally independent training samples containing seven attributes and eight outcomes.[9] The prior probabilities for this sample ranged from 2% to 30%. Chard concluded that a minimum of 200 cases are required for the simple Bayes model and that there was a maximum of approximately 500 cases above which increasing the training sample size would produce no further improvement in the model's accuracy.

Studies similar to Chard's have not been published using conditionally nonindependent training samples. However, Gammerman and Thatcher, in a study with nine outcomes and thirty three attributes, reported that a database containing 4,387 patients was not large enough to permit all relevant combinations of symptoms to be identified in adequate detail so that the proper Bayes model would outperform the simple Bayes model.[10]

(3) Number of Attributes:

Several researchers have reported small differences in accuracy between the different Bayes models with small attribute sets and larger differences in accuracy with larger attribute sets.[4, 9, 10] This led Fryback to question the value of using the proper Bayes model when there was a small attribute set. He hypothesized that with a small attribute set, the contribution of each attribute is stronger than the degradation in performance which results from errors caused by ignoring the dependencies between attributes. Thus, the proper Bayes model is only preferred when there is a large attribute set.

MODEL FACTORS

Model factors influence the relative accuracy of the simple and proper Bayes models through their management of a model's data. The presence or absence of model factors can either enhance or diminish the relative performance of Bayes models.

(1) Attribute Order:

Fryback demonstrated the importance of considering the sequence in which attributes are selected from a conditionally nonindependent training sample for inclusion in proper Bayes models.[4] Common measures of expected attribute impact upon model accuracy which have been used to order model attributes include: information gain, error reduction, and relative informativeness.[11, 12, 13] Attribute order has no significance in the simple Bayes model where there is no reduction in the data set during computations. However, it may have significance in the proper Bayes model if data set reductions prevent attributes from being included in the final model.

(2) Attribute Exclusion:

Ohmann, et. al. reported that none of the conditionally nonindependent models they tested achieved maximum accuracy when all attributes were included.[14] This led them to conclude that better results could be obtained if adequate strategies were used for the selection of attributes.

Ohmann, et. al. identified peaking as a problem for models such as logistic regression and discriminant analysis which partition their training samples and automatically include all attributes. By implication, peaking would be a problem for the proper Bayes model if all attributes were included. In contrast, Ohmann, et. al. also reported that monotonicity could be demonstrated in a simple Bayes model with many attributes and they concluded that it may be preferable to use all attributes in simple Bayes models which do not partition the training sample during computations.

(3) Exclusion Criteria:

Both chi-square and sample theory tests have been used in machine learning systems as measurements for attribute exclusion.[12, 15, 16] However, as Fisher first noted, the real issue may not be the measurement that is used, but rather the confidence level which is chosen for excluding attributes.[17]

Fisher observed a pessimistic bias in machine learning systems that reject an attribute which can not be proven to significantly influence accuracy. He states that pessimistic biases arise because criteria which are important in hypothesis testing are naively translated into criteria which are used for evaluating 'hypothesis plausibility' in learning systems. In contrast to the pessimistic bias, Fisher proposes an optimistic bias. This bias states that an attribute is deemed relevant unless it is demonstrably noninfluential to system accuracy. His research indicates that the choice between optimistic and pessimistic biases in machine learning systems depends upon three factors.

Training Sample Size: Fisher's findings show that optimism, a low confidence level for excluding attributes, is preferable when there is a smaller training sample. After the training sample reaches sufficient size, there is little difference in performance across confidence levels.

Training Signal Noise: Noise is variance in the training data that cannot be modeled by the learning algorithm. Fisher's research shows that optimism achieves better results in training samples with low noise.

Outcomes Distribution: An uneven distribution of prior probabilities creates a pre-existing bias in favor of one of the outcomes. In a training sample where cases were evenly distributed across two outcomes, Fisher found that a pessimistic bias was detrimental to performance, particularly with a small training sample. Conversely, in a training sample where 91% of the observations were in one category, he found that the choice of confidence level had little effect on the outcome.

STUDY METHODOLOGY

This study investigated the impact of variations in training sample size, number of attributes and attribute exclusion criteria upon the accuracy of the simple and proper Bayes models. It was organized as a controlled experiment in which these attributes were manipulated while the remaining situational and model factors described above were controlled. The average area under ROC curves was used to evaluate the discriminatory ability of the models.

Myocardial Infarction Data Base:

All data used in this study were collected under a grant from the Health Care Finance Administration.[18] Sixteen questions in that data base correspond to the factors which are used in the APACHE II system to predict patient outcomes. These sixteen questions were used as the attributes in our models. Another question in that data base describes the patients' discharge status (coded as dead or alive). This question was used as the outcome which the models in this study predicted.[19] The myocardial infarction data base contained 1139 cases which met the APACHE II inclusion criteria.

Training and Test Samples:

Previous studies have compared the performance of different inductive learning systems.[17, 20, 21] The performance criteria most frequently chosen is a system's ability to accurately classify cases which it has not previously seen. In our experiment, the test samples were fixed at a size of 339 cases while their associated training samples were selected in sizes of 100, 400, and 800 cases. These training sample sizes were chosen because they were respectively below, within, and above the 200 to 500 case training sample range which Chard used to measure the simple Bayes model's performance.

Number of Attributes:

Separate from our selection of training samples, we also selected attribute sets. Using a randomization procedure, we selected sets of 4, 8, and 12 attributes from the 16 attributes in the APACHE II data set. These groupings were chosen because they were respectively below, near, and above the seven attributes which were used in Chard's study. We repeated this procedure 30 times for each attribute set size and created a total of 90 different attribute sets for use in this study. Next, we randomly assigned the attribute sets to the previously created training and test samples. This created 10 occurrences for each of the nine training sample size and attribute set size combinations.

B.E.St. Models:

In a previous study, we developed a system which selects its attributes according to their information value and is able to model zero through n-1 orders of conditional nonindependence in Bayes models.[12] This system also may optionally use sample theory as a means of excluding attributes from a Bayes model. This exclusion is effected by using the binomial distribution to test whether an attribute's likelihood ratio is statistically different from unity. If it is not statistically different from one, the attribute is excluded. Kramer and Thieman provide tables which contain the number of training sample cases that are required to detect these differences for a particular likelihood ratio value.[22]

Outcome Prediction:

We used B.E.St to calculate test sample probabilities of patient survival under four scenarios (simple Bayes with optimistic exclusion, simple Bayes with pessimistic exclusion, proper Bayes with optimistic exclusion, and proper Bayes with pessimistic exclusion) for each of the 90 sample-attribute set combinations. Using the probabilities of patient survival, we computed areas under the ROC curve for each of the 360 resulting test samples.[23, 24] These ROC curve areas were then averaged for each of the 36 distinct combinations of training sample size (100, 400, and 800 cases), number of model attributes (4, 8, and 12 attributes), Bayes model (simple or proper Bayes), and attribute exclusion criteria (sample theory or no exclusion).

STUDY RESULTS

Table 1 below shows the main and interaction effects that were identified in our experimental design. Higher order interaction effects, although tested, are excluded from the table as they were not significant. The overall F test value of 0.0001, indicates that the experimental design accounts for a significant amount of the variability in average ROC area. The R-Square value of 57.586% substantiates this finding.

Main Effects:

All four main effects attributes in Table 1 have F tests which are significant at the 0.0001 level. This indicates that the average ROC areas for different values of these attributes are not equal.

Paired Interaction Effects:

Only two interaction effects in our design were significant at or below the .05 level. These were training sample size with exclusion criteria and number of attributes with Bayes model. Two additional pairwise interactions that are cited in the literature (training sample size with Bayes model and number of attributes with exclusion criteria) were not significant. Table 2 shows these relationships.

Table 1:
ANOVA FOR AVERAGE ROC AREA

Pr > F = 0.0001
R-Square = 0.575860

Main Effects	Pr > F
Training Sample Size	0.0001
Number of Attributes	0.0001
Bayes Model	0.0001
Exclusion Criteria	0.0001
Interaction Effects	**Pr > F**
Training Sample Size with Number of Attributes	0.1476
Training Sample Size with Bayes Model	0.3866
Training Sample Size with Exclusion Criteria	0.0001
Number of Attributes with Bayes Model	0.0073
Number of Attributes with Exclusion Criteria	0.6408
Bayes Model with Exclusion Criteria	0.6617

The training sample size with exclusion criteria interaction in Table 2 shows that under optimistic exclusion (no attributes excluded) there is little change in average ROC area (less than 1%) as the number of cases increases from 100 to 800. However, there is a significant change in average ROC area (over 9%) as the number of cases increases under pessimistic exclusion (sample theory is used to exclude attributes).

T tests with an alpha of .05 were used to compare average areas under ROC curves. The results show that with a small or an intermediate number of cases (100 or 400), optimism is preferred to pessimism and with a large number of cases (800), there is no difference in accuracy between the optimistic and pessimistic exclusion criteria. Further, under optimistic exclusion, there is no difference in accuracy between small, intermediate, and large training samples.

With regard to the second significant interaction effect, number of attributes with Bayes model, the simple and the proper Bayes models both increased their accuracy as the number of

Table 2:
AVERAGE INTERACTION ROC AREAS

Situation Factors	Exclusion Criterion	
	Optimism	Pessimism
100 Cases	0.73377	0.63620
400 Cases	0.74238	0.70041
800 Cases	0.74152	0.72832
4 Attribs	0.68934	0.63104
8 Attribs	0.74805	0.70114
12 Attribs	0.78039	0.73275

Situation Factors	Bayes Model	
	Simple	Proper
100 Cases	0.70027	0.66982
400 Cases	0.72801	0.71477
800 Cases	0.74894	0.72090
4 Attribs	0.66125	0.65913
8 Attribs	0.73702	0.71218
12 Attribs	0.77895	0.73418

attributes increased. However, the simple Bayes model increases its accuracy at a greater rate than the proper Bayes model. At 4 attributes, there is no difference between the models, but at 12 attributes simple Bayes is over 4% more accurate than proper Bayes. Thus, with a small number of attributes (4), there is no significant difference in accuracy between the two Bayes models and with a moderate to large number of attributes (8 or 12), the simple Bayes model is preferred.

The two interaction which were are cited in the literature but which were not significant in our study are both special cases. First, optimism is significantly more accurate than pessimism at all attribute levels. Second, simple Bayes is more accurate than proper Bayes with small and large sample sizes. With an intermediate sample size, there is no significant difference in accuracy.

Complex Interaction Effects:

It is often assumed that proper Bayes will outperform simple Bayes when there is a large sample size and a large number of attributes. This interaction was not significant in our design. Simple Bayes was significantly more accurate than proper Bayes with 12 attributes and 100 or 800 cases. In all other situations there was no difference between the two Bayes models.

With regard to exclusion criteria, optimism was more accurate than pessimism with a small or intermediate sized sample, regardless of the number of attributes. With a large sample, there was no difference in accuracy.

DISCUSSION

Situational Factors:

Prior research has proposed that when conditional nonindependence exists, increasing the training sample size will have a greater impact on the accuracy of the proper Bayes model than on the simple Bayes model. Other researcher have proposed that with large training sample sizes, proper Bayes will produce more accurate results than simple Bayes. Our results confirm that proper Bayes became more accurate as the training samples increased in size. However, proper Bayes's accuracy never exceeded that of simple Bayes.

The fact that our simple Bayes models did not become more accurate as more cases were added to the training sample seems to conflict with the results of Chard.[9] This earlier study concluded that a minimum of 200 cases are required for the simple Bayes model and that there was a maximum of approximately 500 cases above which increasing the training sample size would produce no further improvement in the model's accuracy. In contrast, our model achieved a point of maximum accuracy at or below 100 training cases. Despite the seeming contradiction, there are important differences between Chard's study and ours. Chard was concerned with estimating eight outcomes from seven attributes. The prior probabilities for his outcomes ranged from 2% to 30%. Our study estimated one outcome from 4, 8, and 12 attributes and our prior probability was 85%. Thus, with a larger prior probability, we could be expected to achieve more accurate results with a smaller sample.

Several researchers have reported smaller differences in accuracy between the simple and proper Bayes models with a small number of attributes and larger differences in accuracy with larger attribute sets.[4, 14] Our study found that with a small number of attributes, there was little difference between the performance of simple and proper Bayes. However, as more attributes were added, it was the simple, rather than the proper, Bayes model that had the greatest improvement in accuracy.

Manipulating the number of attributes in a simple Bayes model appears to have a far greater impact on model accuracy than is achieved by increasing the training sample size. In fact, it may be more economical in many studies to increase a simple Bayes model's accuracy by increasing the number of attributes that are collected rather than by increasing the training sample size. Again, this relationship should be investigated further.

Model Factors:

Fisher observed that an optimistic bias is preferred when there is a small training sample and that neither bias is preferred with a large training sample. Our results support this observation. With a small or intermediate number of cases (100 or 400), optimistic exclusion produced more accurate estimates. With a larger number of cases (800), there was no significant difference in predictions with either pessimistic or optimistic exclusion. We extended Fisher's work by varying the number of attributes in our models. We found that optimistic exclusion consistently outperformed pessimistic exclusion for all attribute sets (small, intermediate, and large).

When we reviewed the complex interaction between training sample size, number of attributes, and exclusion criteria, we found that optimism is only preferred to pessimism when there is a small or intermediary number of cases (The number of attributes did not matter.). With a large training sample size, there is no difference in performance between optimism and pessimism.

Summary:

Our results tend to support the position that managing conditional nonindependence is not as important a problem as some researchers have assumed and that the simple Bayes model is fully capable of managing conditional dependencies. In our study, simple Bayes consistently outperformed proper Bayes and even proved better in the one situation (large training sample and large number of attributes) where proper Bayes was assumed to excel.

While this study did not discover those situational factors which allowed proper Bayes to outperform simple Bayes, it did eliminate several candidates (sample size, number of attributes, and exclusion criteria). This study also demonstrated that increasing sample size and/or the number of attributes, as has been previously suggested, might not be the best way of handling problems of this type. Instead, other approaches should be investigated which would better manage the data which are available. Perhaps, by selecting attributes for models according to their importance for the particular test case under evaluation instead of their importance in the entire training sample, the accuracy of proper Bayes models could be improved.

REFERENCES

[1] Gustafson, D. H., Cats-Baril, W. L., Alemi, F. *Systems to Support Health Policy Analysis: Theory, Models, and Uses*. Ann Arbor, Michigan. Health Administration Press (1992).

[2] Gustafson, D. H., Kestly, J. J., Greist, J. H., Jensln, N. M. Initial evaluation of a subjective Bayesian diagnostic system. *Health Services Research* 6, 204 (1971).

[3] Norusis, M. J., Jacquez, J. A. Diagnosis. I. Symptom nonindependence in mathematical models for diagnosis. *Computers and Biomedical Research* 8, 156-172 (1975).

[4] Fryback, D. G. Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research* 11, 423-434 (1978).

[5] Seroussi and the ARC & AURC Cooperative Group. Computer-aided diagnosis of acute abdominal pain when taking into account interactions. *Methods of Information in Medicine* 25, 194-198 (1986).

[6] Lichtenstein, S. Conditional non-independence of data in a practical bayesian decision task. *Organizational Behavior and Human Performance* 8, 21-25 (1972).

[7] de Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., Horrocks, J. C. Computer-aided diagnosis of abdominal pain. *British Medical Journal*, II, 9-13 (1972).

[8] Bayes, T. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions* 3, 370-418 (1783). Reproduced in *Two Papers By Bayes*, ed. W. E. Deming. New York, New York. Hafner (1963).

[9] Chard, T. Self-learning for a bayesian knowledge base: how long does it take for the machine to educate itself? *Methods of Information in Medicine* 26, 185-188 (1987).

[10] Gammerman, A., Thatcher, A. R. Bayesian diagnostic probabilities without assuming independence of symptoms. *Methods of Information in Medicine* 30, 15-22 (1992).

[11] Bigongiari, L. R., Preston, D. F., Cook, L., Dwyer, S. J., Fritz, S., Fryback, D. G., Thornbury, J. R. Uncertainty/information as measure of various urographic parameters: An information theory model of diagnosis of renal masses. *Investigative Radiology*, 16, 77-81 (1981).

[12] Alemi, F., Bhatt, P., Eisenstein, E., Fadlalla, F., Stephens, R., Butts, J. Torturing data until they confess: A self learning bayesian expert system (B.E.St.). *Working Paper of the Collaborative Care Project*. Cleveland, Ohio (1992).

[13] Quinlan, J. R. *C4.5: Programs For Machine Learning*. San Mateo, California. Morgan Kaufmann Publishers, Inc. (1993).

[14] Ohmann, C., Qin Yang, Kunneke, M., Stoltzing, H., Thon, K., Lorenz, W. Bayes theorem and conditional dependence of symptoms: different models applied to data of upper gastrointestinal bleeding. *Methods of Information in Medicine*, 27, 73-88 (1988).

[15] Quinlan, R. Induction of decision trees. *Machine Learning*, 1, 81-106 (1986).

[16] Clark, P., Niblett, T. The CN2 induction algorithm. *Machine Learning*, 3, 261-284 (1989).

[17] Fisher, D. *Pessimistic and Optimistic Induction*, Technical Report CS-92-12, Department of Computer Science, Vanderbilt University, Nashville, Tennessee (1992).

[18] Alemi, F., Rice, F., Hankins, R. Predicting in-hospital survival of myocardial infarction. *Medical Care*, 28(9), 762-775 (1990).

[19] Knaus, W. A., Draper, E. A., Wagner, D. P., Zimmerman, J. E. APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13(10), 818 (1985).

[20] Kononenko, I., and Bratko, I. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6, 68-80 (1991).

[21] Shavlik, J. W., Mooney, R. J., Towell, G. G. Symbolic and neural learning algorithms: an experimental comparison. *Machine Learning*, 6, 111-143 (1991).

[22] Kramer, H. C., Thieman, S. *How Many Subjects: Statistical Power Analysis and Research*. Sage Publication, Inc. Newbury Park, California. (1987).

[23] Hanley, J., McNeil, B. The meaning and use of the area under a receiver operating curve. *Diagnostic Radiology*, 143(1), 29-36 (1982).

[24] McNeil, B. J., Hanley, J. Statistical approaches to the analysis of receiver operating characteristic curves. *Journal of Medical Decision Making*, 4(2), 137-150 (1984).